NNPDF: A succesful Machine Learning application to High Energy Physics

Artificial Intelligence Group, University of Cambridge

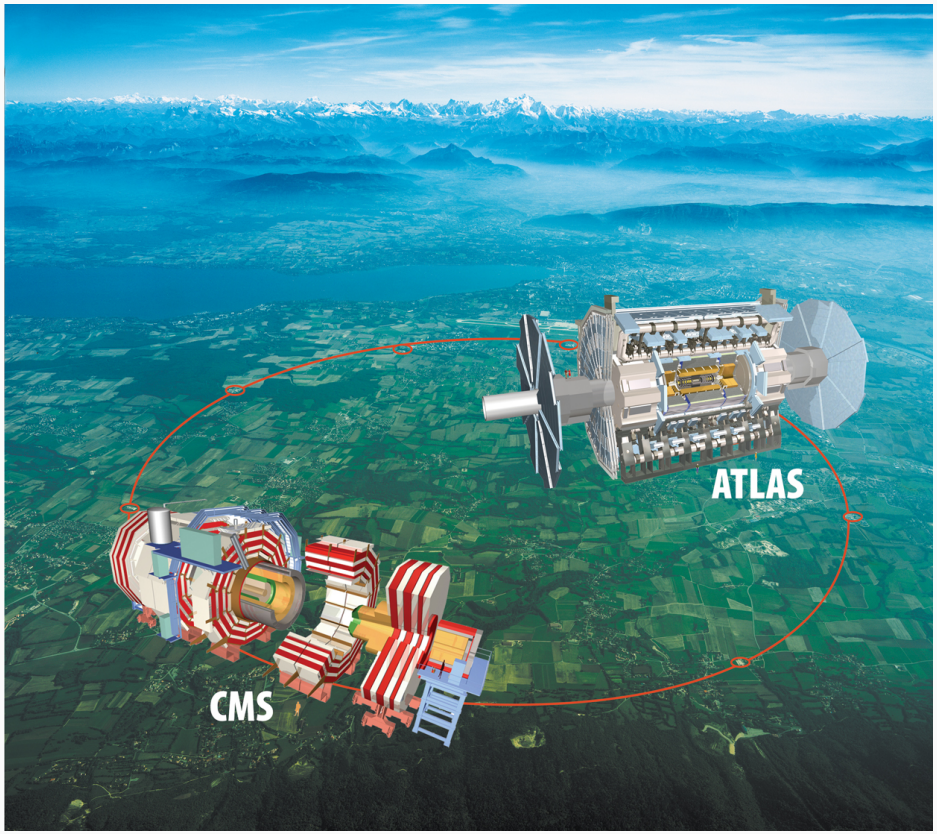
Zahari Kassabov

October 23, 2019

Cavendish Laboratory, University of Cambridge

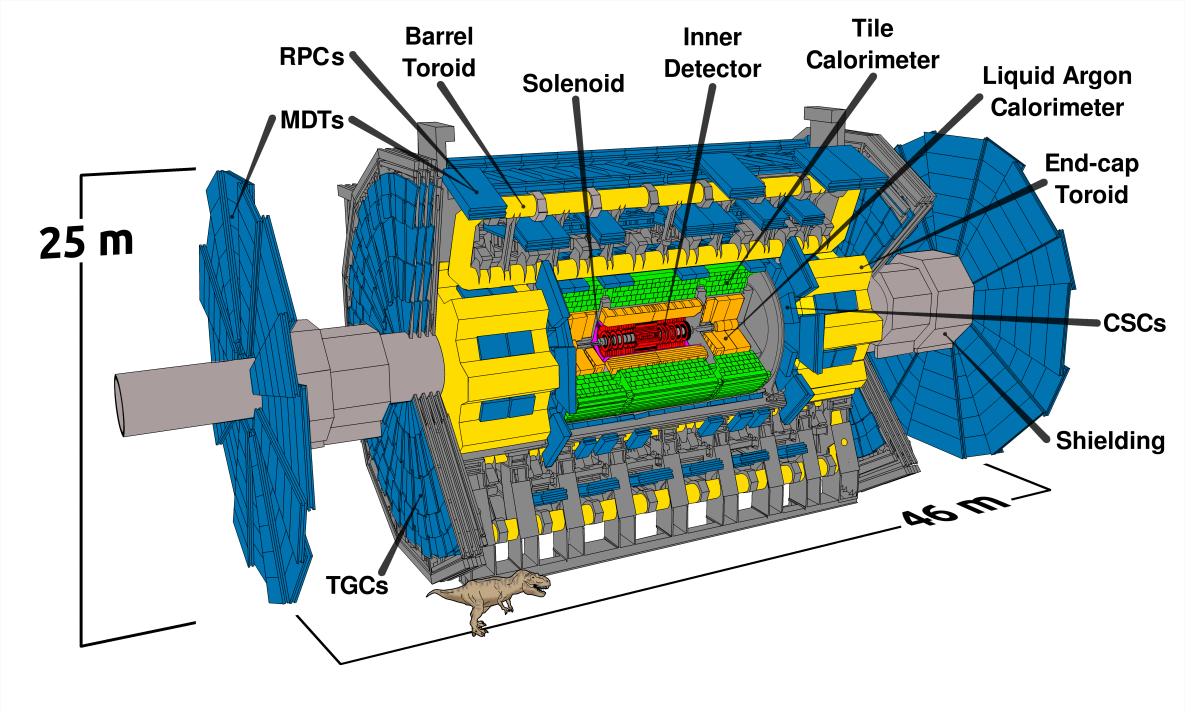
The LHC and CERN

- The Large Hadron Collider at CERN:
  - Collides protons accelerated to ~$(1 - 10^{-9})$ times the speed of light.
  - We learn about Physics by studying the results.
- Lots of opportunities and for IT R&D. Many existing results (www, cloud computing, real time processing..).
- Some interesting applications of ML (this talk).

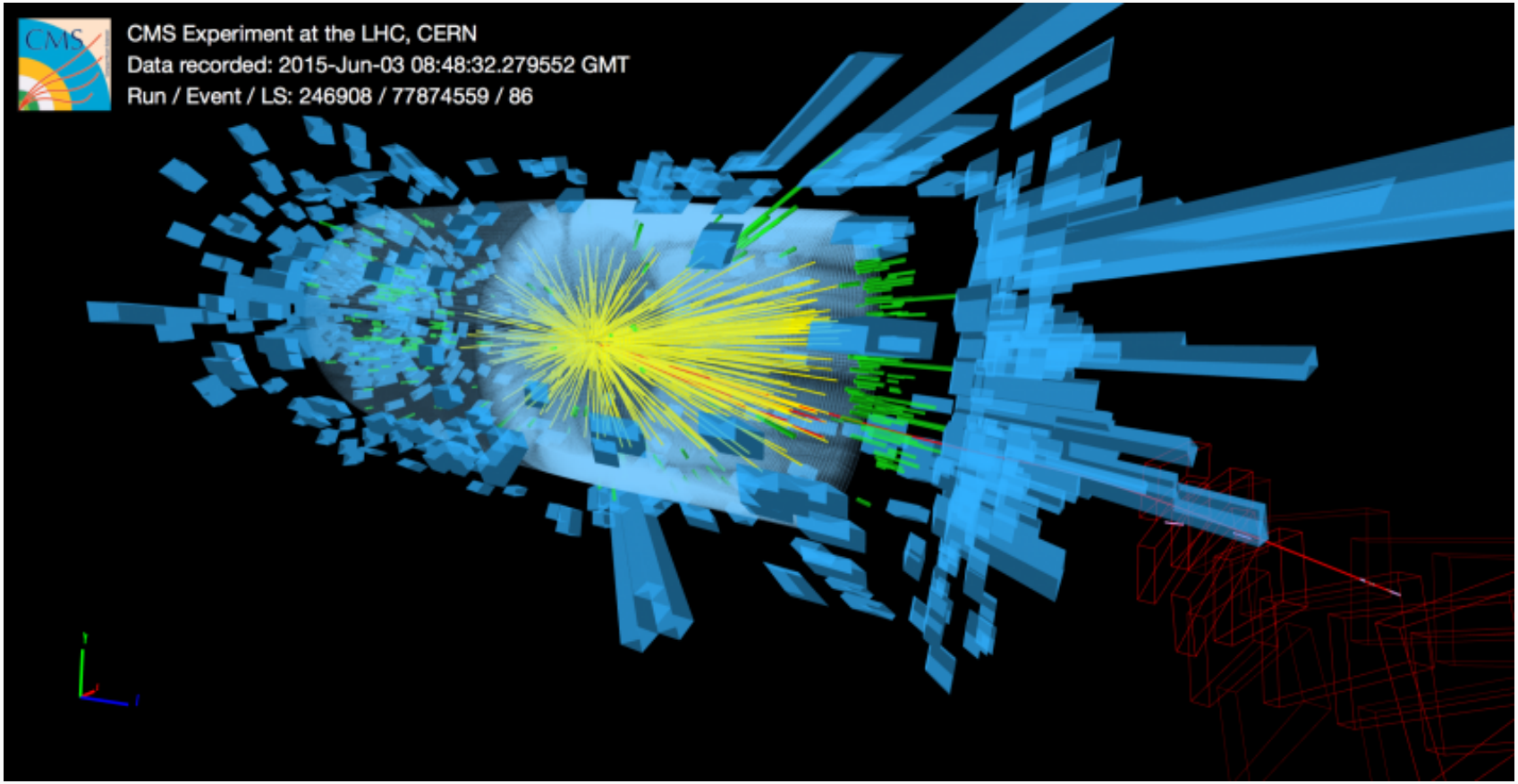


LHC detectors

- Structured in layers specialized in measuring different properties of collision results (momentum of an electron, energy energy depositions,...).

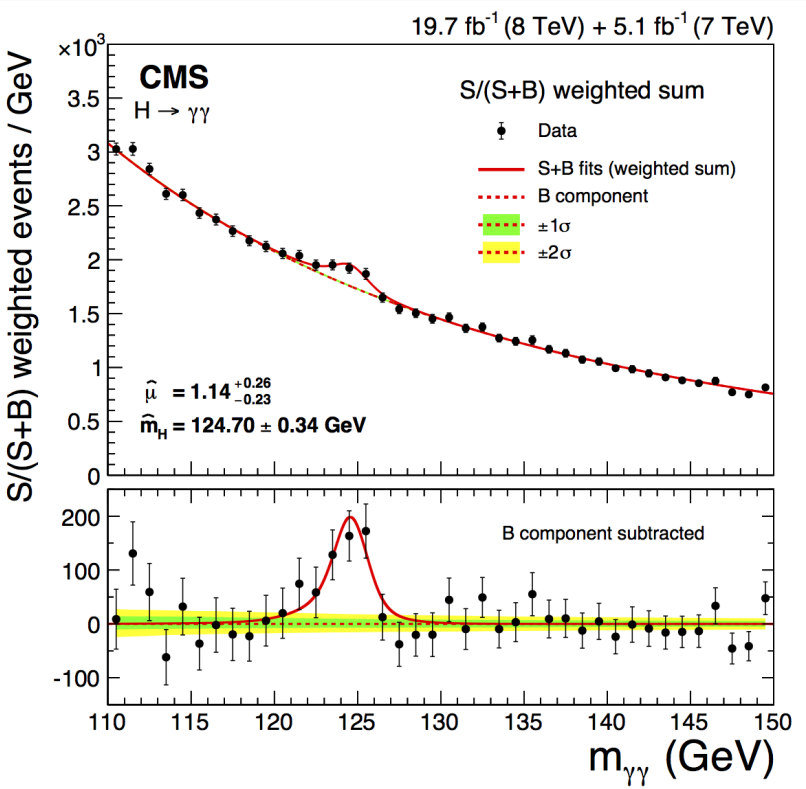


An event display



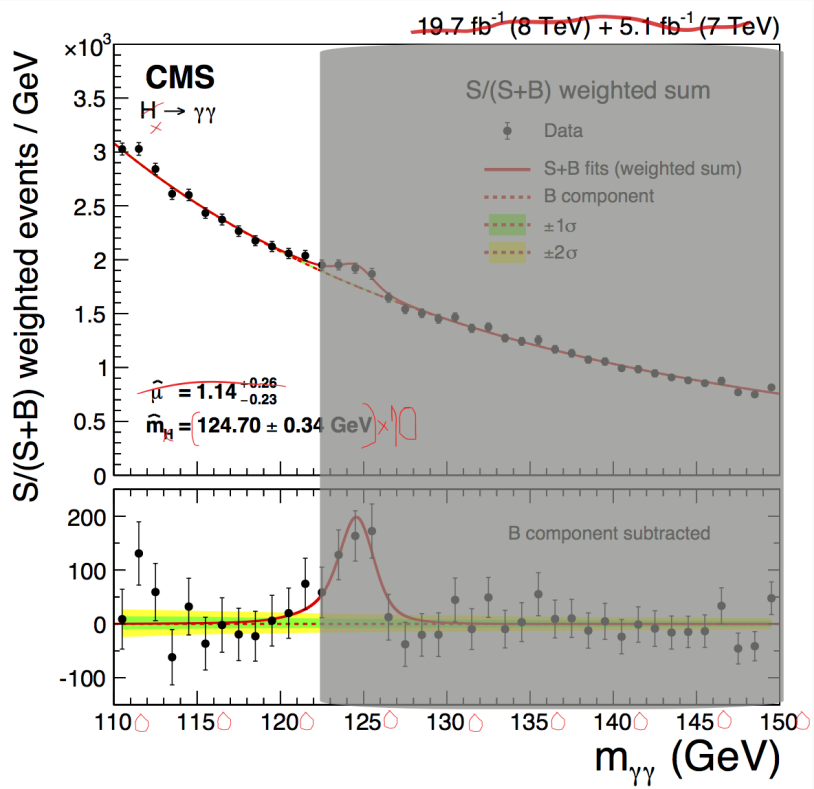
Predictions at the LHC: discovering the Higgs Boson

- Theory predicts a resonance peak in a concrete region of the spectrum of decay products, above the predictions of the “No Higgs” background.



- Great success of theory + experiments (2013 Nobel Prize).
- Theory doesn’t need to be very precise: Only smooth background needed.
- Theory does need to be precise to describe its properties
  e.g. measure spin.

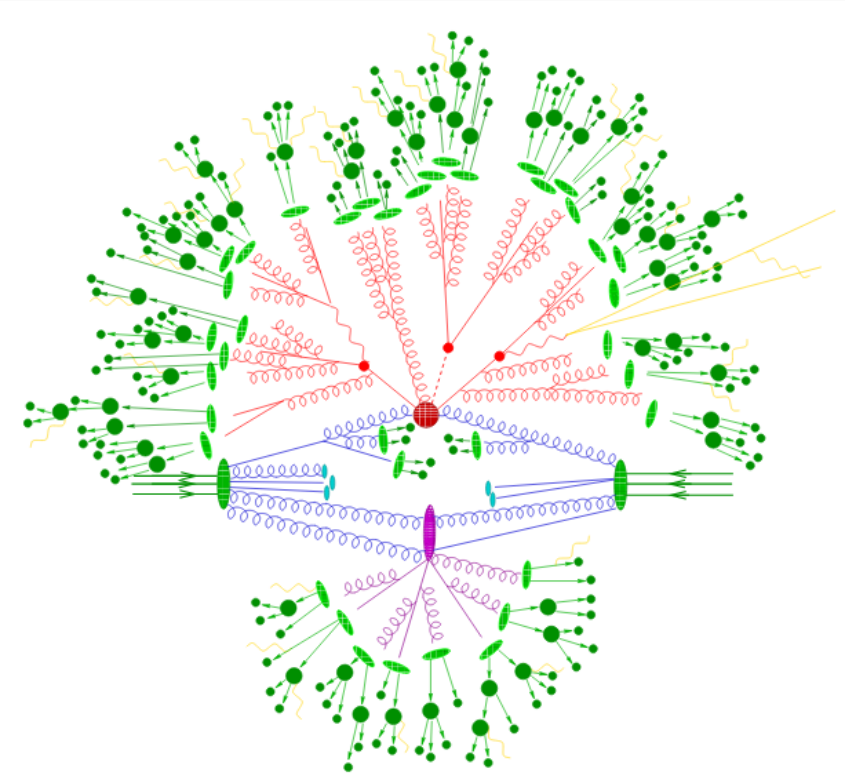
Searching for new Physics



Need extremely precise data and theory!

- LHC set out to find New Physics: Notably an explanation for dark matter.
- So far very successful at *ruling out* direct detection of new particles (i.e. peaks we do not know about) at accessible energy scales.
- Open question: Can we detect new particles at *inaccessible* energies?

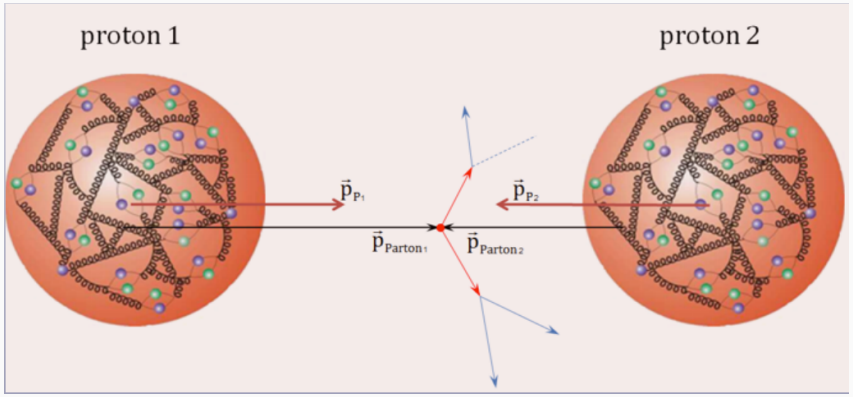
Theory predictions at the LHC



- Need to invert this very complicated system.
- Only high energy interactions can be predicted with fundamental theory (modulo parameters).
- Rest has to be extracted from data.

Parton distribution functions

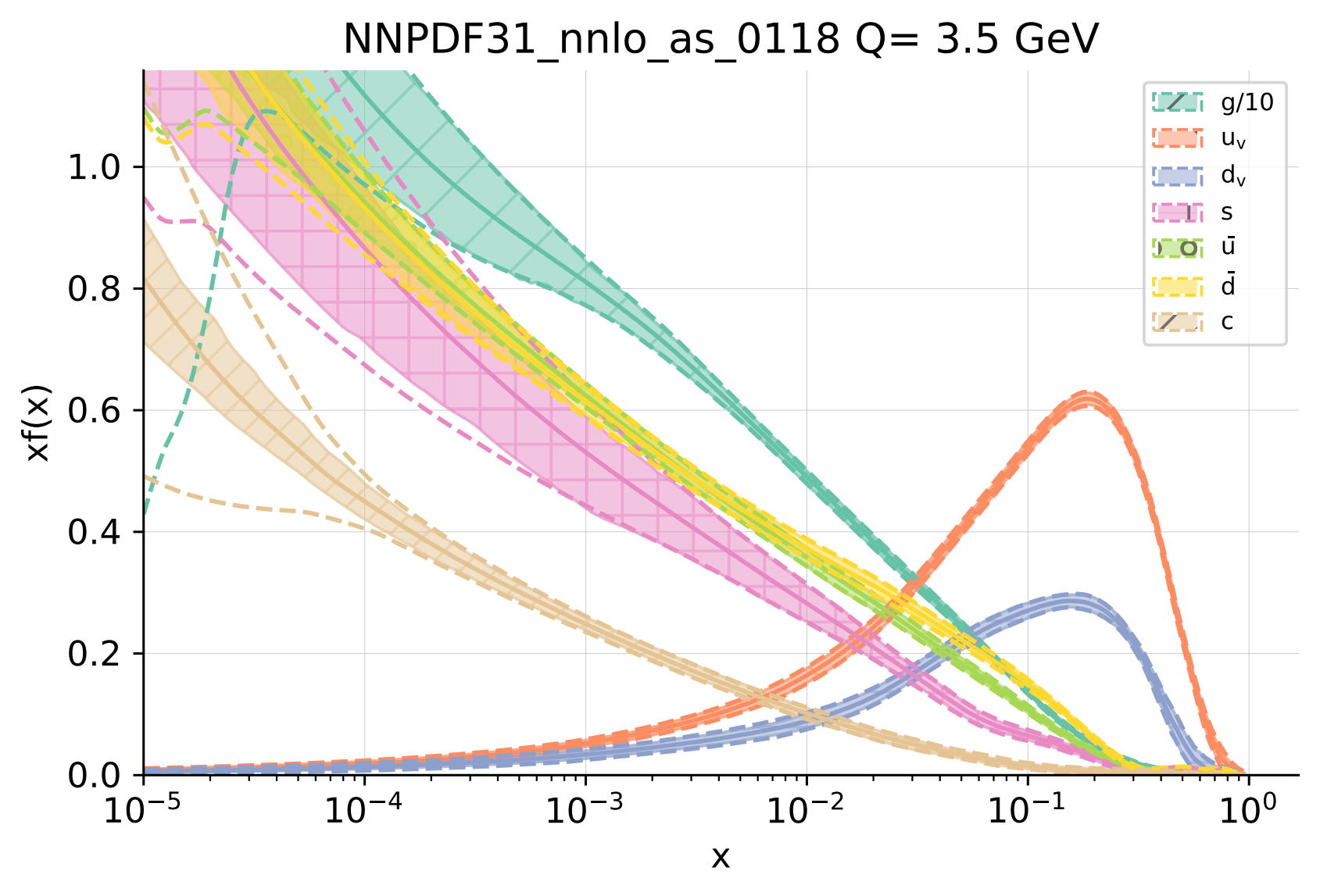
- Roughly speaking, probability distribution of sampling a *parton* (e.g. quark or gluon) from a proton, in a high energy collision, with a given momentum.



- Variables:
  - Longitudinal momentum of parton ($x = p_{\text{parton}}/p_{\text{proton}}$).
  - Energy of the (partonic) collision.

E.g. $u(x = 0.1, Q = 3.5 \text{ GeV})dx$ means probability of sampling quark up with fraction of momentum between 0.1 and $0.1 + dx$ in a 3.5 GeV collision.

Some PDFs


NNPDF31_nnlo_as_0118 Q= 3.5 GeV
g/10
u_v
d_v
s
$\bar{u}$
$\bar{d}$
c
xf(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}

What we know about PDFs

Not much about x dependence, from first principles. Only "sum rules".

- Conservation of momentum

$$\sum_{i}^{\text{partons}} \int_{x=0}^{x=1} x f_i(x, Q) dx = 1$$

- Quantum valence numbers:

$$\int_{x=0}^{x=1} (u(x, Q) - \bar{u}(x, Q)) dx = 2$$

$$\int_{x=0}^{x=1} (d(x, Q) - \bar{d}(x, Q)) dx = 1$$

$$\int_{x=0}^{x=1} (f_i(x, Q) - \bar{f}_i(x, Q)) dx = 0 \;\; i \in \{s, c, b, t\}$$

Q dependence completely determined from theory (renormalization group equation).

- ML problem reduced to finding all of the $f_i(x)$ at some fixed scale Q_0.

Determining PDFs

Determining PDFs is:

- Important: Dominant source of uncertainty in many important analyses.
- Challenging: Effectively 8 “functional degrees of freedom”.
- Can be done by relating things we know how to calculate (*partonic cross sections*) with things we know how to measure (*hadronic cross sections*).

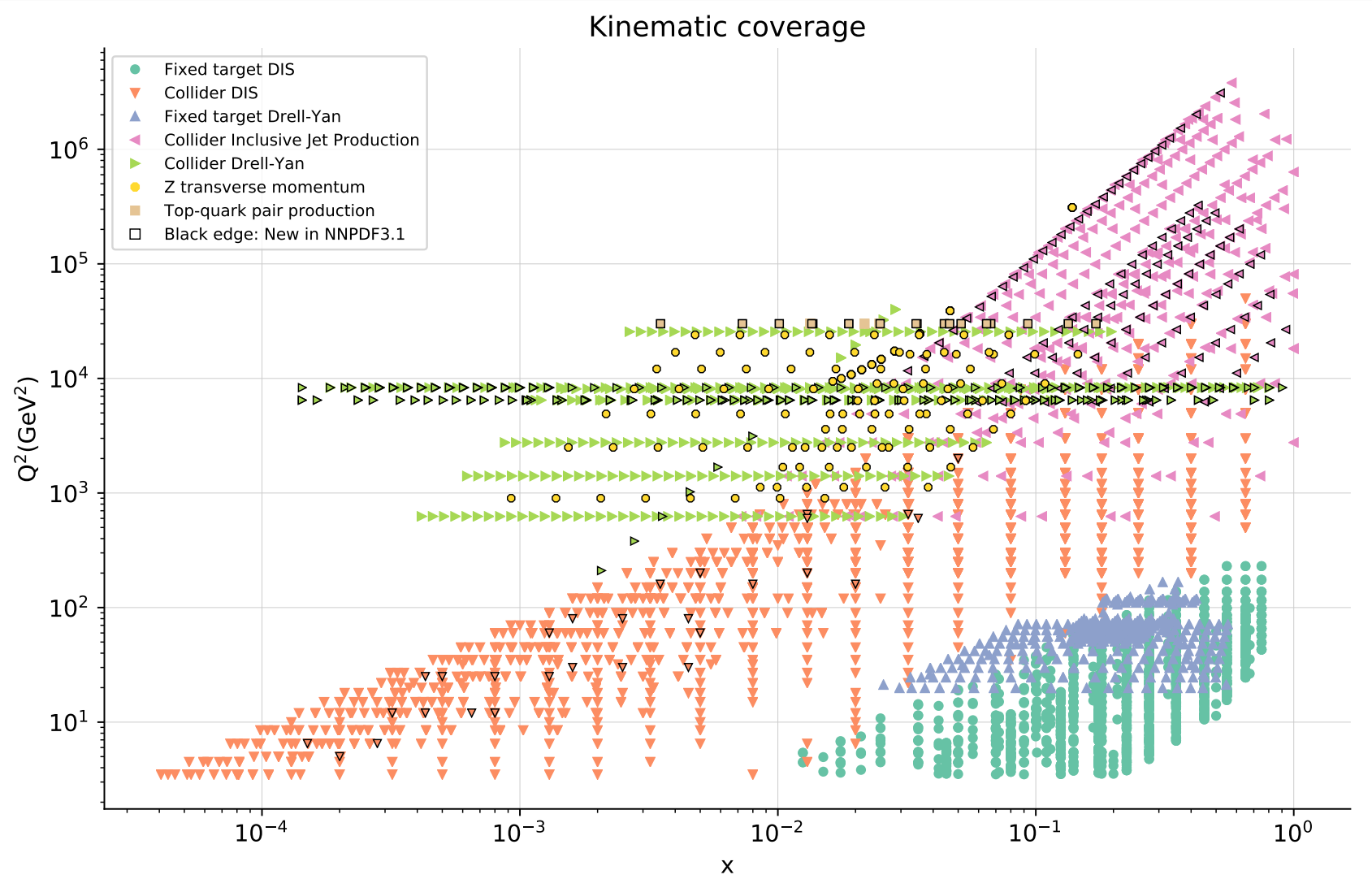
$$\sigma_{pp\to X} = \sum_{i,j}^{\text{partons}} \int_0^1 \int_0^1 dx_1 dx_2 \sigma_{ij\to X}(x_1, x_2, Q) f_i(x_1, Q) f_j(x_2, Q)$$

Experimental data

- The input for the fit is a collection of hadronic cross sections, corresponding to various kinematical regions and processes.
  - Roughly 4000 points, so not “big data”.
  - However note complicated relation between PDFs and input data.
- With good approximation, data has Gaussian Uncertainties, but with non trivial correlations of experimental uncertainties.
  - Can see it as a distribution with mean given by the experimentally measured central values and the experimental covariance matrix Σ.
- Our loss function is the maximum likelihood estimator,

$$\chi^2 = \sum_i^N \sum_j^N (\text{data}_i - \text{prediction}_i)\Sigma_{ij}^{-1}(\text{data}_j - \text{prediction}_j) = \delta^T \Sigma^{-1} \delta$$

Our dataset


Kinematic coverage
Fixed target DIS
Collider DIS
Fixed target Drell-Yan
Collider Inclusive Jet Production
Collider Drell-Yan
Z transverse momentum
Top-quark pair production
Black edge: New in NNPDF3.1
Q^2(GeV^2)
10^6
10^5
10^4
10^3
10^2
10^1
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
x

What we need

Three ingredients required to determine PDFs (from the ML point of view):

- A way to parametrice the functions.
- A way to fit the parameters to data.
- A way to propagate uncertainties.
  - From the data.
  - From interpolation and extrapolation.

PDF fits: The old way.

- Assume a simple functional form (weakly motivated by theory considerations) and fit the parameters to data:

$$f(x) = Ax^{\alpha}(1-x)^{\beta}$$

- Add more parameters ad-hoc when it doesn't fit, e.g.:

$$f(x) = Ax^{\alpha}P(x)(1-x)^{\beta}$$

  with P(x) some polynomial.
- Propagate experimental uncertainties with linear error propagation.
- In general uncertainties too optmistic as the model choice itself is not considered a source of uncertainty.

Enter NNPDF

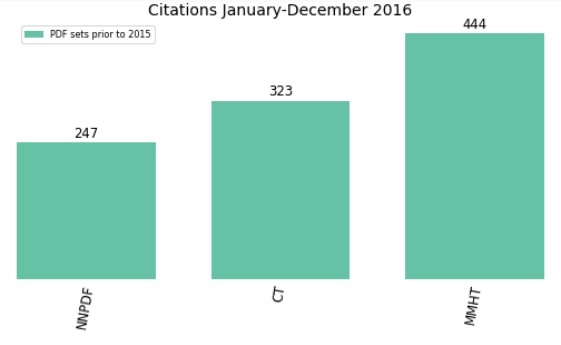
Key ideas (circa 2002).

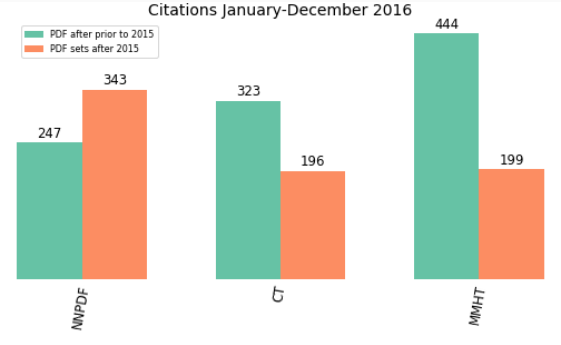
- Use neural networks to parametrice PDFs. Avoid “theoretical bias” of selecting a restricted model.

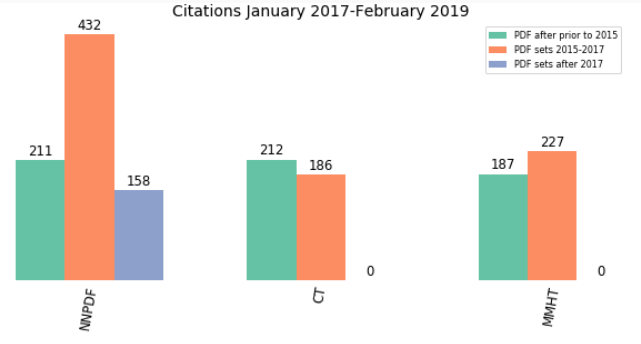
$$f(x) = Ax^{\alpha} NN(x)(1-x)^{\beta}$$

- Use a “Monte Carlo replicas” to propagate uncertainties and deliver the result.

How popular was that?







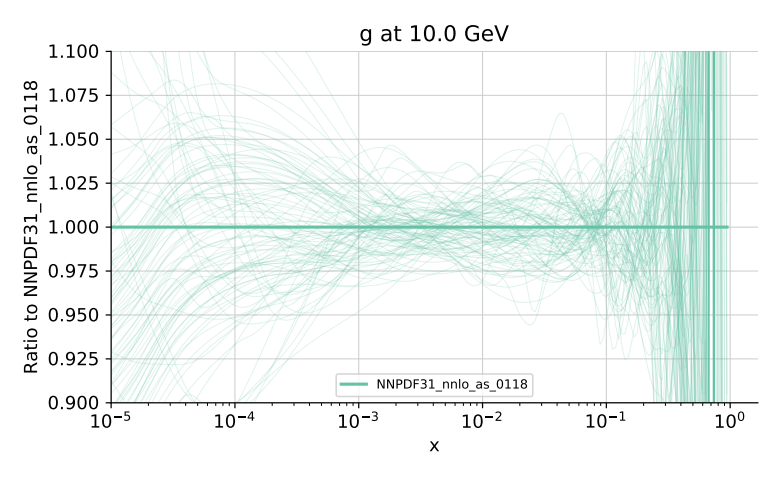
Monte Carlo replica sampling

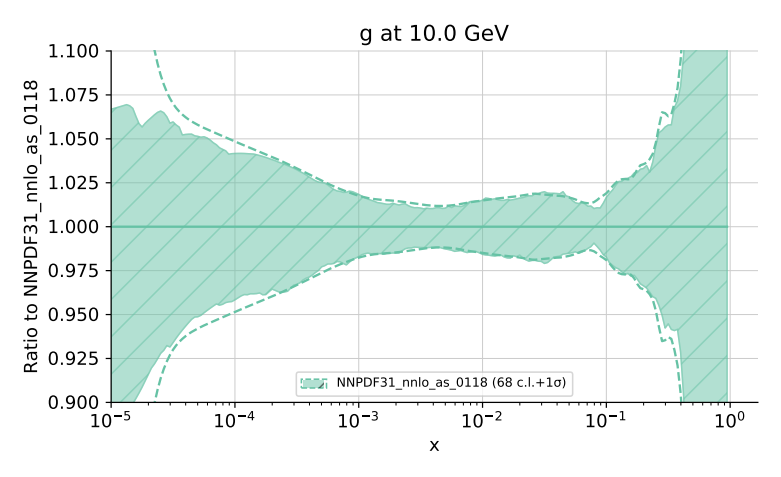
- Experimental data has uncertainties. Generally optimization depends on some random state. We want to propagate these uncertainties to the final result.
- Idea: sample "*pseudo datasets*" from the distribution of experimental data and produce a different fit for each (*replica*).

$$d_j^{(k)} = d_j + \Sigma_{ij}^{1/2} n_j^{(k)} \ , \ n_j^{(k)} \sim \mathcal{N}(0,1)$$

- To get the *PDF uncertainty* of a prediction, compute it for each replica and look at the resulting sample.

PDF replicas



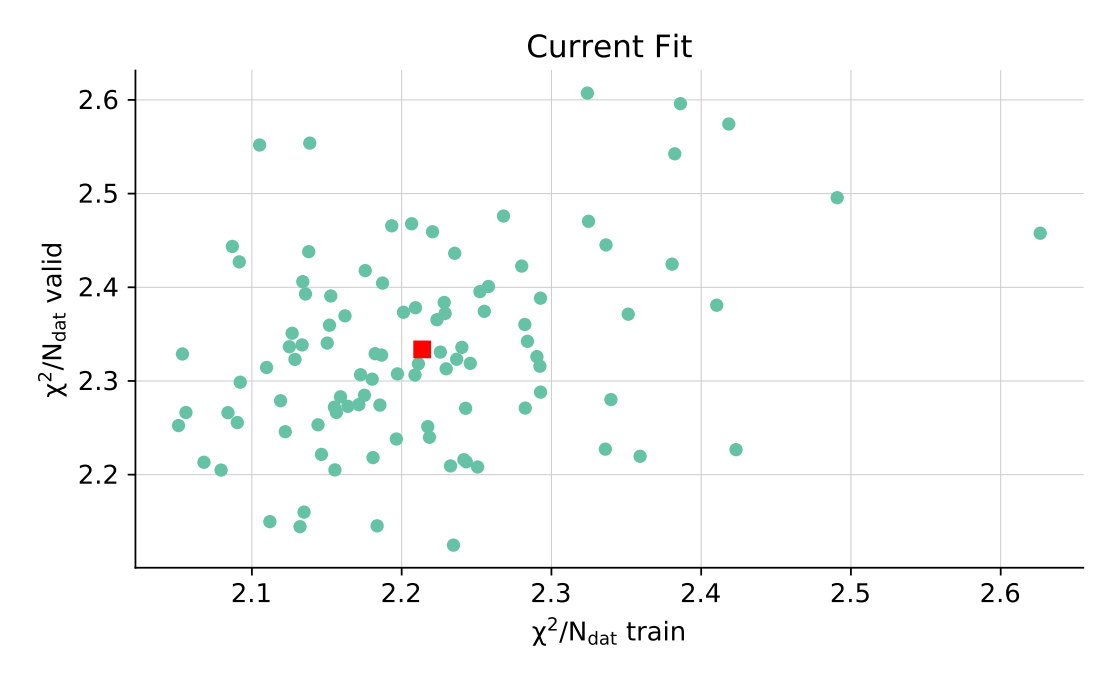


Optimization methodology

- Currently large scale effort to revamp the whole methodology. Will discuss a mixture of current and new (experimental) results.
- Main difference so far: New optimization based on gradient descent with Keras + Tensorflow. Old optimization based on genetic algorithm on a custom C++ code.
- Factor ~10 improvement in training time. Can use extra efficiency for hyperoptimization (but not completely obvious what do we want to optimize for..).

Cross validation and stopping

- For each replica, we split each dataset in half: a training and validation subset.
- We optimize on the training subset but select the configuration that has the best validation error function.
- For GA train for a fixed number of iterations and select best. For GD stop when validation stops improving.

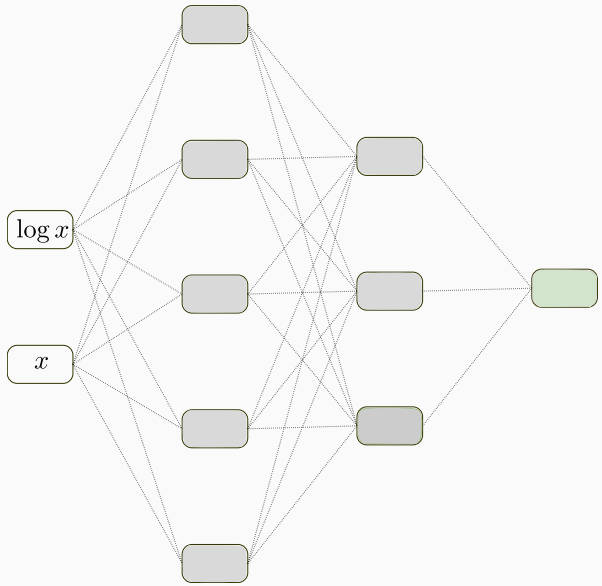


Genetic algorithm

- Full details at https://arxiv.org/pdf/1410.8849.pdf#subsection.3.3
- At each iteration, create “mutants” by fluctuating parameters in the current best (i.e. best training error function) result.
- Select best mutant for next iteration.
- Heavily hand tuned. Significative more wiggly replicas, but fit quality similar to GD.

PDF parametrization

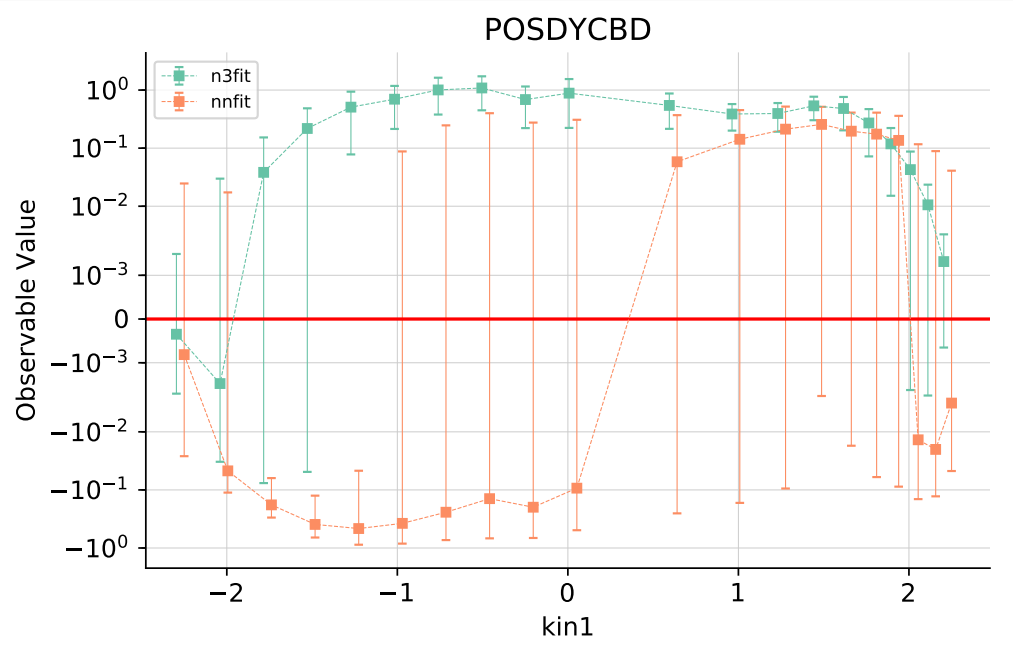
- We have $f(x) = Ax^{\alpha} NN(x)(1-x)^{\beta}$
- $NN(x)$: simple feed-forward multilayer perceptrons (sigmoid hidden layers and linear output layer).



- Exponents random and fixed in GA fits, with iterated ranges. Just fitted in TF.
- Now experimenting with a single network with multiple outputs for all the PDFs and other architectures.

Positivity

- Contrary to usual probability density functions, PDFs do not need to be positive everywhere (because partonic cross section isn't).
- But hadronic cross sections have to be positive.
- We impose positivity by adding theory predictions for which we have no data and requiring they are positive (by adding some penalty term to the error function).
- In practice very challenging to add these positivity constraints.



Post selection of replicas

In practice some replicas do not converge to reasonable results at the end of the fit. We need to impose some cuts.

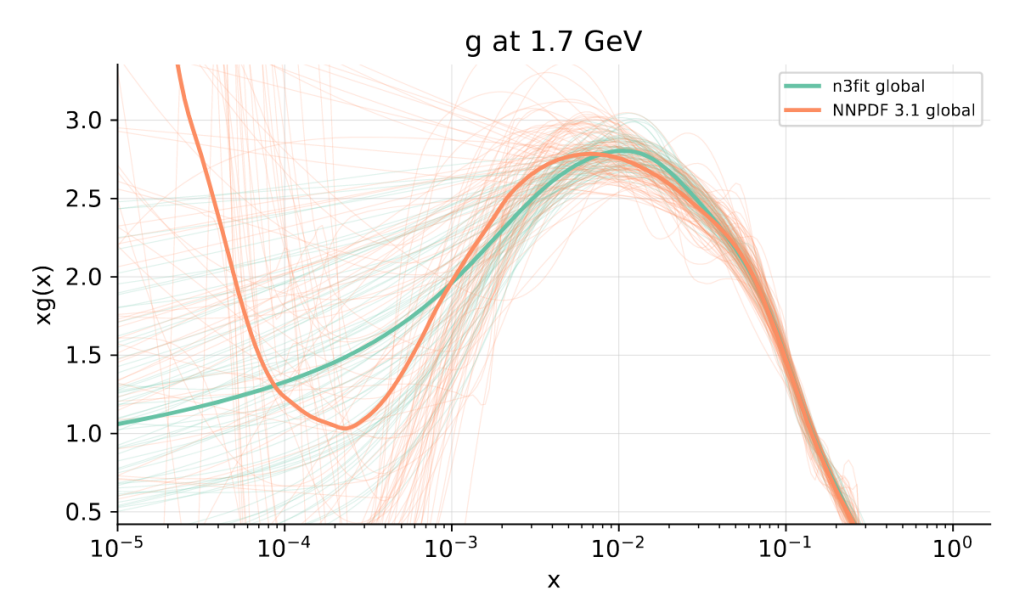
- Positivity could not be fitted.
- χ^2 too high (compared to other replicas).
- Replica too wiggly (high arc-length).

We have to impose veto on these quantities. For χ^2 and arc length, roughly 5 sigma if these were distributed Gaussianly.

- Future hyperoptimization should seek to minimize these outliers.

Extrapolation

- Regions at very large or small x not constrained by available data or theory. Yet crucial for high energy predictions.
  - Currently GA and GD give very different results.



- Is there a principled way to decide which extrapolation is good? Bayesian methods such as Gaussian Processes?

Bad total χ^2

- If the experimental data are sampled from the model that we obtain, and we applied appropriate cross validation, the χ^2 statistic should follow a χ^2 distribution, so we expect $\chi^2 = N \pm \sqrt{2N}$.
- Yet we find a significatively (at $\sim 4\sigma$) higher value. Possible explanations:
  - Problems in the theory (we know it is approximate).
  - Incorrect fixed theoretical parameters (e.g. strong coupling constant).
  - Problems in the experimental data.

How do we know our result is correct?

- People have been using it to make precise comparison of data and theory and found good agreement.
- Closure tests:
    1. Replace experimental central values (but not uncertainties) by the theoretical predictions with the PDFs of our competitors.
    2. Find that we get qualitatively the same PDF as the starting one (they agree within uncertainties).
    - Working on making this more quantitative.

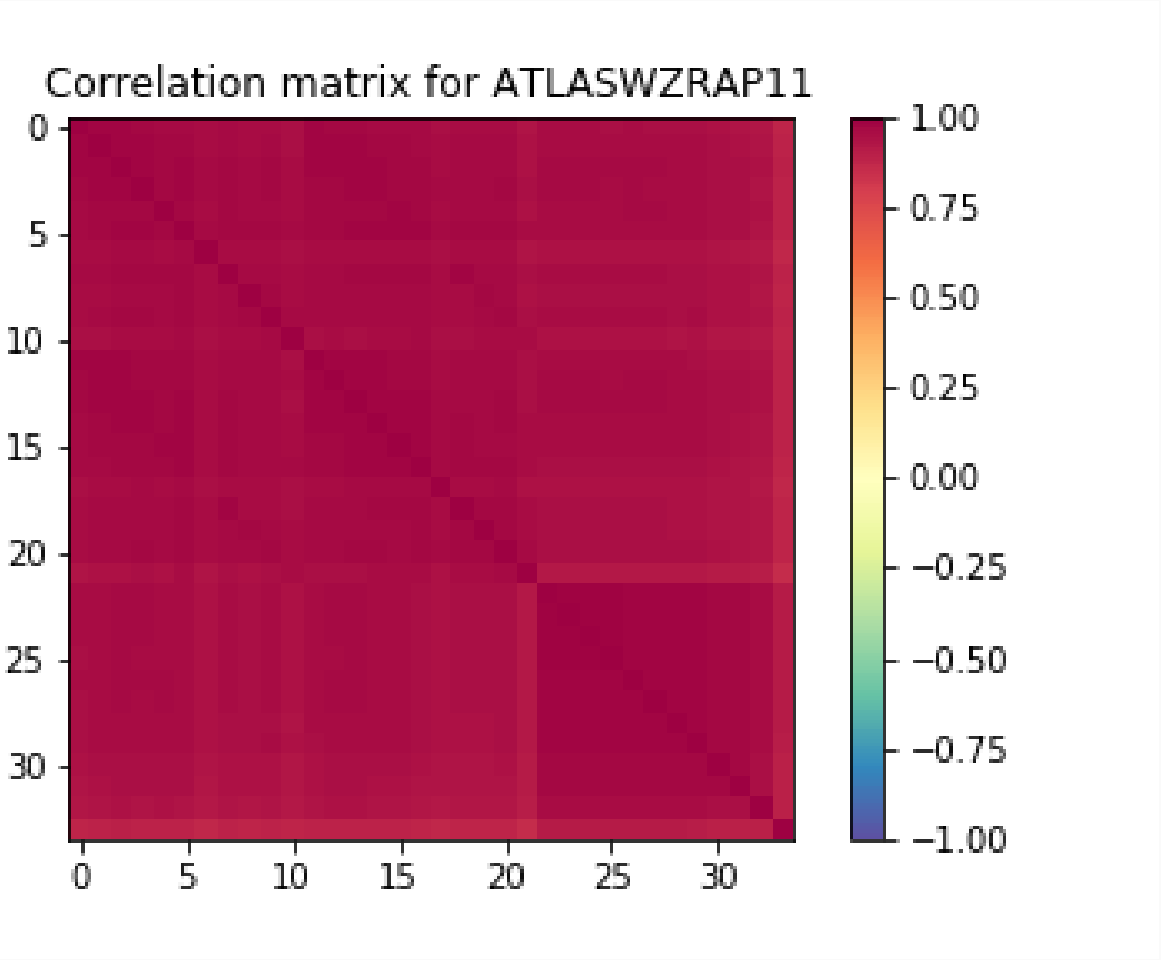
Unstable experimental covariance matrices

- Correlations of experimental uncertainties (off diagonal entries of the covariance matrix) not estimated very precisely.
- Covariance matrices frequently close to singular.
- Uncertainty propagates amplified to the χ^2 when inverting the covariance matrix.
- Some evidence that appropriate regularization might be enough to bring the χ^2 to the expected value.

Regularization strategy

- Idea: Large contributions to the χ^2 should not come from components of the covariance matrix that are not constrained precisely enough.
- Take the square root correlation matrix.
- Clip the smallest singular values to some threshold.
    - Don't yet know how to find the threshold in general..
- Reconstruct a new covariance matrix.

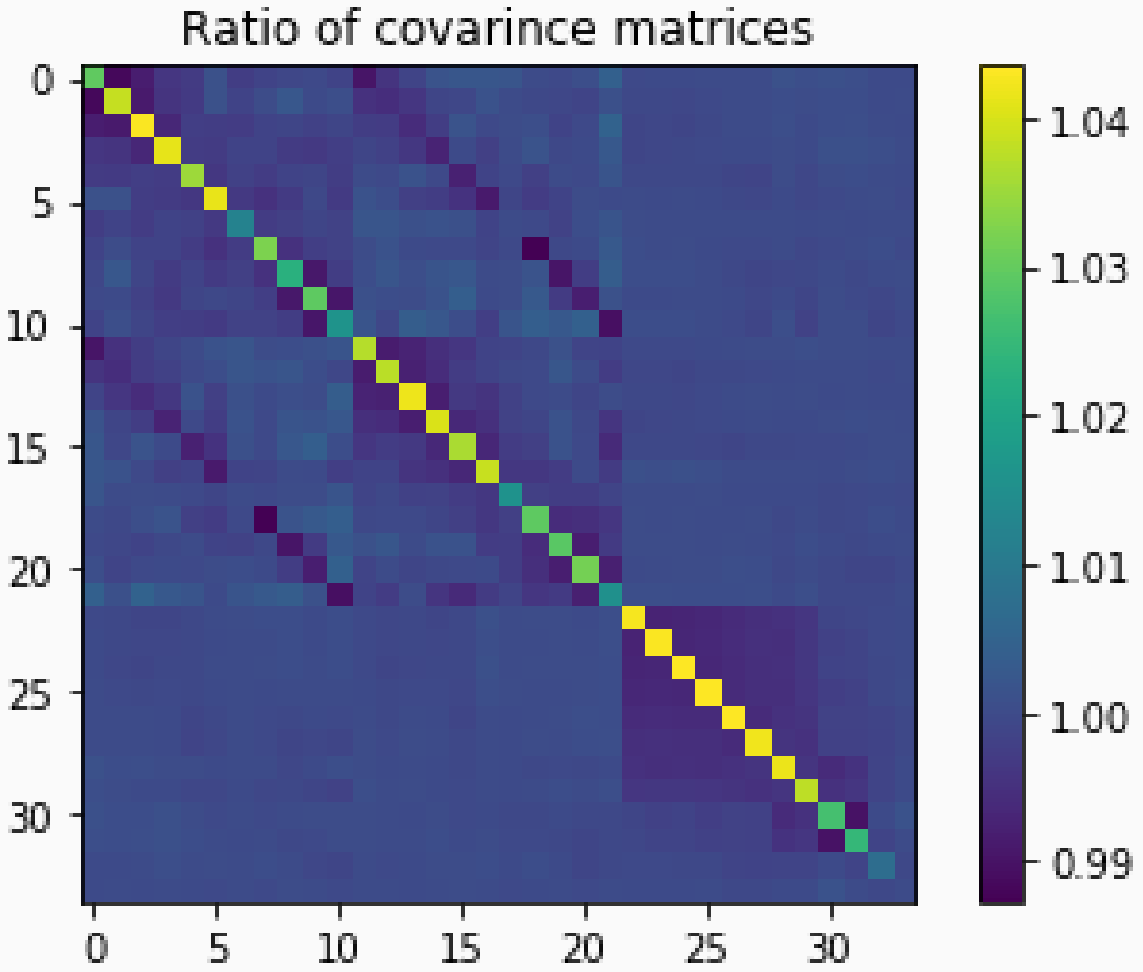
Example



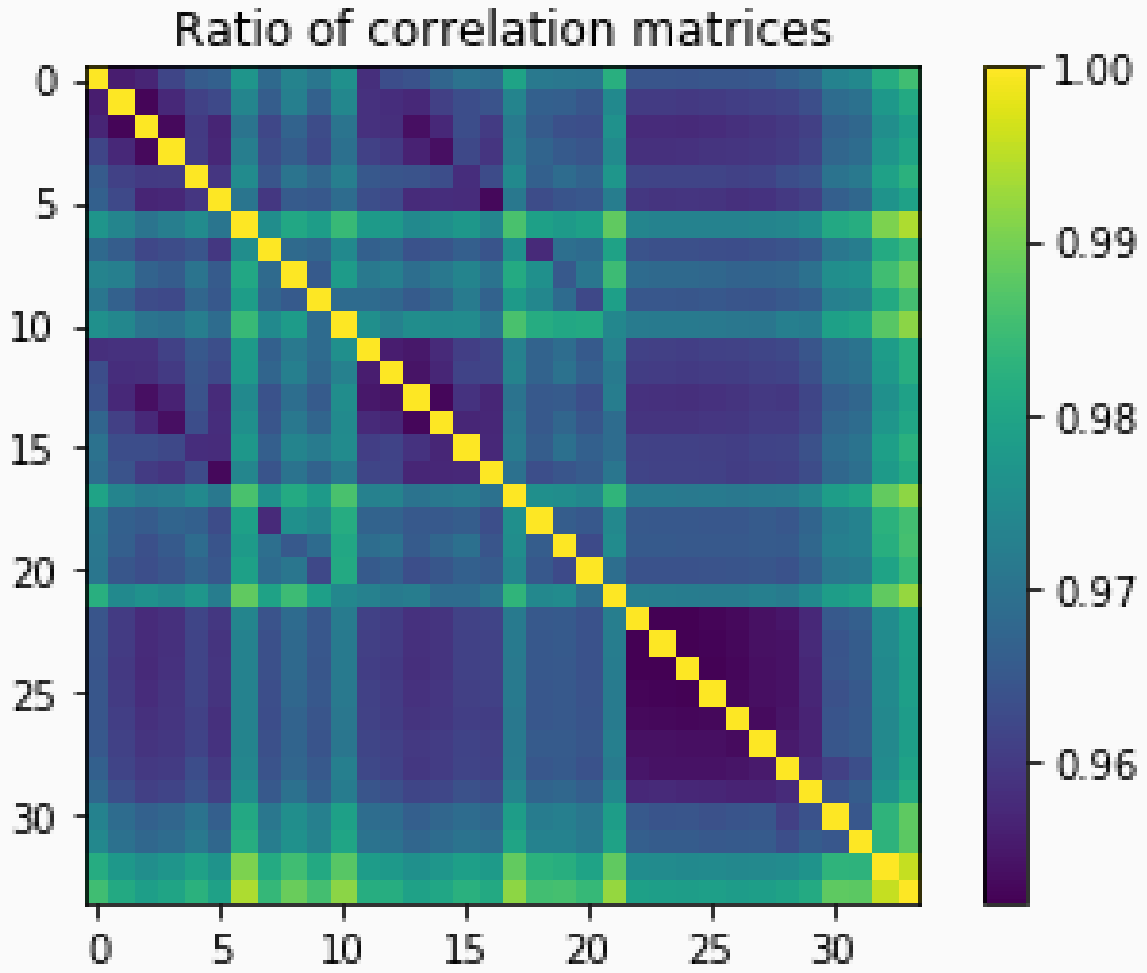
$$\chi^2/N = 71/34$$

Let's apply the regularization..

Covariance matrix ratio



Correlation matrices ratio



Final χ^2

After regularizing and refitting..

$$\chi^2/N = 40/34$$

Minuscule change in the covariance matrix (below the experimental precision) causes large correction in the χ^2.

Thank you!